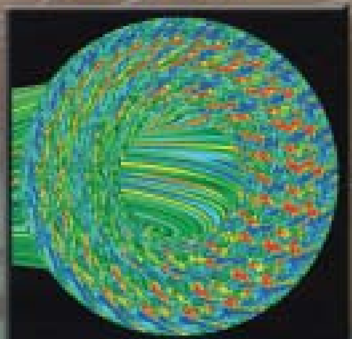
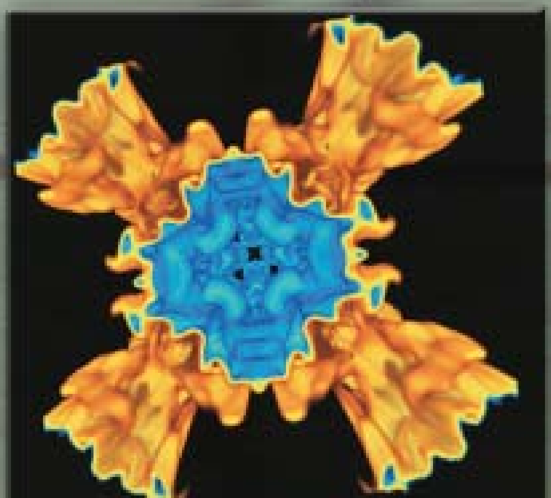
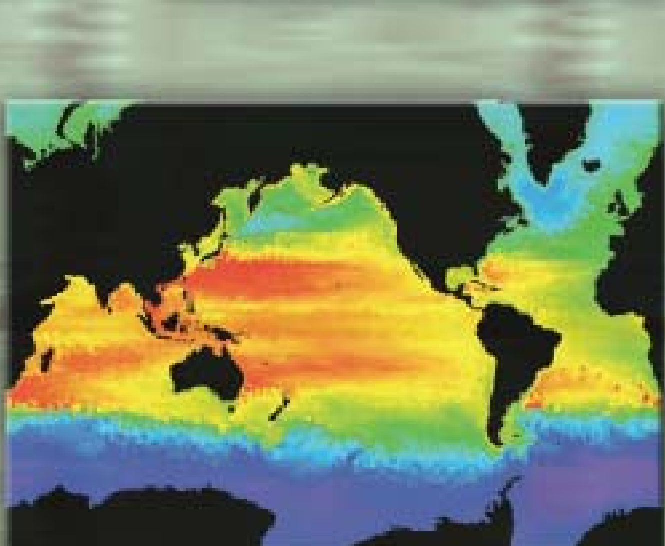
DOE Science Networking Challenge: Roadmap to 2008







Office of Science
U.S. Department of Energy

DOE Science Networking Challenge: Roadmap to 2008

Report of the June 3-5, 2003, DOE Science Networking Workshop
Conducted by the Energy Sciences Network Steering Committee at the request of the Office of Advanced Scientific Computing Research of the U.S. Department of Energy Office of Science

Workshop Chair

Roy Whitney

Report Editors

Roy Whitney
Larry Price

Energy Sciences Network Steering Committee

Larry Price, Chair
Charles Catlett
Greg Chartrand
Al Geist
Martin Greenwald
James Leighton
Raymond McCord
Richard Mount
Jeff Nichols
T.P. Straatsma
Alan Turnbull
Chip Watson
William Wing
Nestor Zaluzec

Working Group Chairs

Wu-chun Feng
William Johnston
Nagi Rao
David Schissel
Vicky White
Dean Williams

Workshop Support

Sandra Klepec
Edward May

Argonne National Laboratory, with facilities in the states of Illinois and Idaho, is owned by the United States Government and operated by The University of Chicago under the provisions of a contract with the Department of Energy.

Available electronically at http://www.doe.gov/bridge and at
http://www.es.net/hypertext/welcome/pr/aboutesnet.html
#programs

Available for processing fee to U.S. Department of Energy
and its contractors, in paper, from:

U.S. Department of Energy
Office of Scientific and Technical Information
P.O. Box 62
Oak Ridge, TN 37831-0062
Phone: (865) 576-8401
Fax: (865) 576-5728
Email: reports@adonis.osti.gov

We are grateful for the editorial assistance of Steve Corneliussen at Jefferson Lab. Argonne's Information and Publishing Division prepared the document for publication, and we are thankful for the services of Vic Comello who served as editor, Judi Damm who developed the document's graphical design, and Gary Weidner who managed print production.

Cover design by Judi Damm, Information and Publishing Division,
Argonne National Laboratory

Table of Contents

Tables

Figures

DOE Science Networking Challenge: Roadmap to 2008

U.S. Department of Energy (DOE) Science Networking resources are crucial for achieving DOE's 21st century science mission. This report puts forth a roadmap for the networks and collaborative tools that the Science Networking and Services environment requires for DOE science fields including astronomy/astrophysics, chemistry, climate, environmental and molecular sciences, fusion, materials science, nuclear physics, and particle physics. Success in these fields depends on scientists' ability to move large amounts of data, access experimental and computing resources via the network, and collaborate in real time from multiple locations across the country and around the world. Implementation of the roadmap presented in this report will be a critical element in keeping DOE a leader in world-class scientific discoveries.

Across the globe, new networking capabilities are emerging and being enthusiastically incorporated — examples include computational and data grids (large numbers of computers and data resources working together across networks), high-speed wireless networking, super-high-speed metro-scale networks for relatively nearby sites, all-optical routers and switches, and inexpensive connections to local computers. Each new capability enables substantial new collaborative functions and efficiencies. However, sophisticated structures and services can be used effectively only if the network infrastructure itself provides the necessary environment. Increasingly, the network must become a collaborative medium for exchanging information, with a core of higher-level services supported by the network providers, in addition to meeting the basic requirements of bandwidth and connectivity. Thus, this report, the result of workshop input from 66 of the nation's leading scientists and their collaborators, proposes initiatives in three areas:

- **Production and high-impact networking.** The operational "production" services that defined the early generations of scientific networking must continue to evolve. Also needed are "high-impact" network services for high-rate transfers of increasingly enormous volumes of data — the terabytes (millions of megabytes) and even petabytes (billions of megabytes) that at present can be handled and analyzed only at originating locations.
- **Technology, services, and collaboratory tools.** Emerging from R&D programs are new, higher-level capabilities in the areas of collaboratory tools and middleware, the software that makes disparate software applications "interoperable," much as the World Wide Web does, and "manageable" as a system of facilities spread nationally, and globally. These emerging capabilities need to be operationally supported more broadly, with systematic progression from R&D to pilot programs to long-term production use.
- **Network research.** A separate, dedicated, R&D network is needed to allow the testing of new protocols while permitting science to proceed in parallel without interruptions caused by network failures and by test requirements for extremely high bandwidth.

The goal of updating DOE Science Networking aligns directly with national priorities as articulated in June 2003 by the directors of the Office of Science and Technology Policy and the Office of Management and Budget.[1] Without an enriched information infrastructure supporting DOE science, fewer breakthroughs would be accomplished and fewer answers to research questions would be obtained with the available funds. For DOE to achieve the goals of its investments in new scientific directions, DOE networking and services must match or exceed the worldwide pace of development.

New costs for the proposed effort start at an estimated \$15.5M in Year 1 and grow, as more capabilities are incorporated to \$21.5M in Year 5. Since the FY 2003 budget for ESnet, middleware, collaboratory pilot programs, and network research is \$39M, the increased funding for the new DOE Science Networking and Services capabilities amount to a 55% growth by the end of the 5-year period.

[1] *Memo M-03-15 from John H. Marburger III and Mitchell E. Daniels, Jr., for the Heads of Executive Departments and Agencies, dated June 5, 2003.*

1. Introduction

This report establishes a roadmap for a new approach to the DOE Science Networking and Services needed for science in the U.S. Department of Energy in the 21st century. It has become increasingly clear[2] that the network provided for DOE science in the past will not be adequate to keep that science competitive in the future. This roadmap, if implemented and followed during the next five years, will solve that problem. The past 5 years have seen a broad and general movement toward the assumption of and reliance on networked systems in all of the large new initiatives for DOE science. It is clear that the success of science depends increasingly on the ability of scientists to move large amounts of data, access computing and data resources, and collaborate in real time from multiple remote locations. It is also abundantly clear that business-as-usual in the network and information services that underpin the scientific collaborations will fall woefully short of what is needed. New capabilities such as computational and data grids, high-speed wireless networking, super-high-speed metro-scale networks, and cheap gigabit Ethernet have arrived in turn and have been enthusiastically incorporated into the arsenal of science, each permitting substantial new collaborative abilities and efficiencies. However, sophisticated structures and services using basic network connections can be used effectively only if the network infrastructure itself provides the necessary environment. Increasingly, the network must become a collaborative information exchange, with a core of higher-level services supported by network providers in addition to basic bandwidth and connectivity.

The August 2002 workshop, High-Performance Networks for High Impact Science, and its report[2] studied in some detail the network requirements of the coming generation of science programs and facilities in the DOE Office of Science (SC), using scenarios submitted by investigators in each of the SC programs. Analysis of these scenarios led to these conclusions (quoting from the report):

- Increasingly, science depends critically on high-performance network infrastructure, where much of science already is a distributed endeavor or rapidly is becoming so.
- We can define a common "infrastructure" with advanced network and middleware capabilities needed for distributed science.
- Paradigm shifts resulting from increasing the scale and productivity of science depend on an integrated advanced infrastructure that is substantially beyond what we have today.

  These paradigm shifts are not speculative. Several areas of DOE science already push the existing infrastructure to its limits as they implement elements of these approaches. Examples include high-energy physics with its worldwide collaborations distributing and analyzing petabytes of data; systems biology access to hundreds of sequencing, annotation, proteome, and imaging databases that are growing rapidly in size and number; and the astronomy and astrophysics community that is federating huge observation databases so it can, for the first time, look at all of its observations simultaneously. The clear message from the science application areas is that *the revolutionary shifts in the variety and effectiveness of how science is done can only arise from a well integrated, widely deployed, and highly capable distributed computing and data infrastructure, and not just any one element of it.*

It is no accident that these observations and the urgent need to update the science information infrastructure fit remarkably with the national priorities for science and technology articulated by the directors of the Office of Science and Technology Policy and the Office of Management and Budget in their memo of June 5, 2003[3] with the subject "FY 2005 Interagency Research and Development Priorities." That memo says, in part, "In general, the Administration will favor investments in Federal R&D programs that sustain and nurture America's science and technology enterprise through the pursuit of … critical research fields *and their enabling infrastructure*"

[2] *http://doecollaboratory.pnl.gov/meetings/hpnpw/finalreport/*

(emphasis added). Of the five Interagency Priorities for R&D Budgets listed in the memo, one is *Networking and Information Technology R&D*.

This memo's focus on networking recognizes advanced networking infrastructure as a basic enabler of present-day science, and as an area that presents great opportunities for the future empowerment of modern science and technology. The rapid pace of advances in the world of networks and network services, as well as the specific interest of the broader federal government, present both a challenge and an opportunity for DOE science. It is time to make a concerted effort to systematically embrace the rush of new capabilities and to formulate a detailed plan to use them to keep DOE science at the forefront of a new generation of scientific discoveries.

The August 2002 High-Performance Networks for High Impact Science workshop report called for the development of a roadmap for an integrated infrastructure that would include:

- A new network provisioning model supporting an integrated three-element network with *production-level networking* in support of traditional program requirements; *network resources for high-impact DOE science programs*, including science application and grid research; and *network resources for network research* that enable experimentation with new concepts.
- Enabling middleware research.
- Enabling network research and accelerating the deployment of the fruits of this research in the service of science.
- A network governance model appropriate to these integrated functions.

The June 3–5, 2003, workshop that led to the present report brought together a broad group of 66 experts, including active investigators from DOE science programs and experts on network operations and emerging network capabilities. They represented universities, national and international laboratories, Internet 2, National Lambda Rail, USAWaves, and three major U.S. telecommunications vendors, plus the DOE Office of Science itself. These participants with their various backgrounds and in some cases competing interests agreed with the following key points:

- The roadmap for production, high-impact, and research networks being presented in this report is the most effective and efficient path for the Office of Science to achieve its networking-related scientific goals,
- The Office of Science networking requirements differ significantly from standard commercial IP requirements and university requirements,
- The production and high-impact network boundary is at the 10 Gbps (i.e., lambda) level for the foreseeable future,
- Only the Office of Science would do much of the network research necessary to meet the Office of Science requirements in a useful time frame,
- Collaborating with university, international, and commercial partners where possible would be very beneficial,
- Central management of the production and high-impact networks with a centrally managed collaboration for the research network would prove to be the most cost-effective and efficient way to achieve the Office of Science networking requirements, and
- Doing the R&D and then providing the core services to support collaboratory tools including grid technologies is critical to the ongoing efficient and effective infrastructure support of DOE science.

This workshop was one of a series of workshops orchestrated by several agencies with goals associated with advancing science. Appendix H lists and describes several of these influential earlier and June 2003 related workshops and conferences. This workshop started from the requirements of the August 2002 High-Performance

[3] *Memo M-03-15 from John H. Marburger III and Mitchell E. Daniels, Jr., for the Heads of Executive Departments and Agencies, dated June 5, 2003.*

Networks for High-Impact Science workshop report and developed a detailed roadmap listing milestones and estimated costs for providing the DOE Science Networking needs to keep DOE science competitive during the next five years. The infrastructure for science was planned in the following categories (which are detailed in the indicated later sections of the present report):

- Section 3 provides a plan for the production and high-impact network pieces of the three-part provisioning model.
- Section 4 identifies 13 middleware technologies and services in priority order and provides a detailed plan for deployment. The top five technologies were judged to be essential, and the next 3 as very important for the support of DOE science. Appendix D provides detailed descriptions and roadmaps for these 8 technologies and services.
- Section 5 outlines a coordinated plan of network research and the network resources needed for developing and testing the new capabilities in a way that can be coordinated with the production and high-impact network functions for efficiency.
- Sections 6 and 7 map structures for management and governance issues identified by the previous workshop.

Section 2 of the present report describes the rapidly evolving overall context for this roadmap. The result of implementing the roadmap during the next 5 years will be the substantially more capable, flexible, and cost-effective DOE Science Networking that will enable DOE science programs to make the most productive use of their research funding. If DOE does not take advantage of this opportunity to support its science with an enriched information infrastructure, less science — in other words, fewer breakthroughs and fewer questions answered — will be accomplished with the available funds. Since the European Union and individual European countries, including the UK and The Netherlands, are making plans for a substantial expansion of networking in support of research and education, we can expect that corresponding support in the U.S. will be needed to maintain the strong record of U.S. leadership in science. DOE runs the risk of negating its investment in new scientific directions if it does not provide correspondingly sophisticated infrastructure.

The new capabilities needed to meet the challenge posed by DOE science programs require a somewhat higher level of investment in the information exchange infrastructure. This investment is needed, however, to enable the effective use of the much larger investments being made directly in the science programs themselves. The costs of the additional capabilities are summarized in Section 8.

As indicated in the Executive Summary, new costs start at an estimated $15.5M in Year 1 and grow, as more capabilities are incorporated into the operating and supported networks, to $21.5M in Year 5, the last year considered in this 5-year roadmap. Since the FY 2003 budget for ESnet, middleware, collaboratory pilot programs and network research is $39M, the increased funding for the new DOE Science Networking capabilities amount to a 40% growth in the first year and a 55% growth by the end of the 5-year period. These increases are justified, considering how the enhanced networking and services infrastructure would be beneficial to the potential for scientific discovery across the Office of Science.

2. Achieving DOE Science Yesterday, Today, and Tomorrow

The mission of the DOE Office of Science is to achieve scientific discoveries by the most effective means possible with an optimal use of resources. Advanced supercomputers and experimental facilities play a vital role in pushing the frontiers of scientific discovery. Accordingly, the Office of Science funds 10 world-class laboratories and a large number of research groups at universities, and collaborating institutions. In this system, the three most valuable resources are:

- Highly trained collaborative groups of scientists having a wide spectrum of talents and backgrounds;
- World-class scientific tools, many of which are at the billion dollar scale of investment of federal resources; and
- Infrastructure and management systems that enable the scientists to make effective use of the tools.

The system is inherently large and complex. Scientists and engineers with diverse backgrounds frequently form both small and large collaborations to make scientific discoveries by taking advantage of various resources and adapting the tools and systems so as to make them an integral part of their daily working lives. They continuously work to improve their scientific tools and systems so that they can advance science.

One of the most useful advancements for science over the last half century has been the rapid evolution of integrated circuit technology. For the past several decades, the density of components on an integrated circuit has doubled every 18 months, and this trend is expected to continue unabated into the next decade. This growth rate, known as Moore's Law [1], has been incorporated in the technology roadmap of the global semiconductor industry [2]. For science, the impact of this increasing capability in processing power lies in increasingly more evolved and complex experiments performed faster and at much larger scales. Two corollaries are (1) that the amount of data that is produced is also rapidly increasing, and (2) the scientific environment is becoming more collaborative and complex. The first challenge has been dealt with by the rapid evolution of computing and networking infrastructures. In fact, networking capabilities have increased faster than Moore's Law for two decades. The second challenge has been dealt with by the evolution of collaboratory/middleware tools, such as the World Wide Web [3], which was invented in a high-energy physics laboratory to improve sharing of experimental data and information.

Science-driven networking requirements for achieving discoveries derive from three factors:

- The volume of data, both experimental and from simulations;
- The collaborative tools used for analyzing and understanding the data; and
- The visualization, computational steering, and other desktop computing tools used by scientists.

Advances in all three of these areas have resulted in the growth of traffic on the Office of Science's Energy Sciences Network (ESnet), which has doubled every year since 1992. To fully appreciate this, understand that on any single day today, ESnet transports more bits of information than it did for the entire years of 1992 and 1993 combined! To help in understanding the scientific drivers, the following table provides some specific examples of DOE scientific goals and the associated experimental, simulation, and analysis data going to media that are involved in achieving the goals. Much of this information is from the August 13-15, 2002, workshop report, *High-Performance Networks for High-Impact Science.*

Table 2-1 Science Data Network and Collaboratory Drivers

1995 – 1999	2002 – 2004	2007 – 2009
Climate In 1998, there were about 5 TB/year of experimental and simulation climate data going to media. About this time, the DOE and other agencies launched a long-range program to acquire experimental data and support simulations.	Climate experimental data and modeling data at the three largest U.S. facilities currently totals 100 TB (NERSC – 40 TB, ORNL – 40 TB, and NCAR [non-DOE] – 20 TB) and is being added to at a rate of 20 TB/year.	By 2008, network-assessable climate experimental and simulation data in the U.S. will be increasing at rate of 3 PB/year. This is due to greatly enhanced experimental measurements and simulations.
Fusion Energy Plasma physics/fusion research at DOE's three main experimental facilities — General Atomics, MIT, and PPPL — and numerical simulations generated 2 TB of data in 1998 (mostly from experiments).	Present plasma physics/fusion experiments and simulations are generating 20 TB/year of data (each contributing roughly half).	Driven mainly by large-scale advanced simulations and preparation for a burning plasma experiment, fusion researchers will be generating 1 PB/year of data by 2008. They also need the necessary collaborative tools to be full partners in the international program.
Hadron Structure Investigation of the quark-gluon structure of the nucleon and nuclei resulted in 50 TB of data and analysis the first full year of operation of all of the experimental facilities of CEBAF at JLab in 1998.	Currently CEBAF experiments and analysis, including those associated with the discovery of the pentaquark, produce 300 TB/year of data.	CEBAF's upgrade to 12 GeV to investigate quark confinement and detailed quark distributions will produce several PB/year.
Quark-Gluon Plasma The goal for the RHIC at BNL is discovering the quark-gluon plasma thought to exist at the edge of the Big Bang. RHIC began operations in 2000.	RHIC has early results that indicate that it may have discovered the quark-gluon plasma and is currently putting 600 TB/year to media.	By 2008, RHIC will increase the amount of data going to media to 5 PB/year as it details its information on the quark-gluon plasma.
Materials Science – Neutrons Neutron Science is critical for investigating the properties of materials by neutron scattering.	The SNS is currently under construction at ORNL. It will increase the U.S.'s neutron science capabilities by more than an order of magnitude.	The SNS will turn on in late 2006 and achieve full operation in 2008, at which time it will produce 200 TB/year of data and analysis.
Materials Science – Photons The four DOE-funded light sources (ALS, APS, NLS and SSRL) are used to investigate the properties of materials and the structure of biological molecules, such as proteins. In 1998, they accumulated 3 TB of data.	Currently the four light sources are acquiring and sending data at the rate of 30 TB/year over ESnet.	The drive to understand the dynamics as well as the structure of materials and biological molecules using greatly enhanced detectors will result in at least a 5-fold increase in the acquisition of data at the light sources by 2008 to 150 TB/year.

1995 – 1999	2002 – 2004	2007 – 2009
Chemistry – Combustion Simulations for combustion are critical to improve our use of energy. The simulations were generating 100 GB/year in 1998.	Construction of a Web-based archive for collaborative sharing and annotation of a broad range of chemical science data is now under way. Combustion is currently generating 3 TB/year and is storing annotated feature and data subsets to this archive.	In 2007, combustion simulations will produce several PB/year of data to be collaboratively visualized, mined, and analyzed. In addition, there will be several 100s of TB/year of experimental data generated, plus publication and annotation in Web-accessible archives of 100s TB/year for collaborative research.
Chemistry – Environmental EMSL at PNNL came on-line in 1997 with the mission of understanding and controlling the molecular processes that underlie our environmental problems. In 1998, it put 250 GB to media.	EMSL's unique combination of simulations, high-field magnetic resonance instruments, high-performance mass spectrometers, optical imaging instruments, and more generate 100 TB/year to media.	As high rate proteomic and nanoscale facilities and high-end supercomputers come on-line, EMSL's rate of putting data to media will increase to 2 PB/year by 2008.
Genomes to Life In the area of proteomics and metabolomics for Genomes to Life (GTL), there was less then 10 GB of data on-line in the world in 1998.	Proteomics and metabolomics currently are capable of generating 400 TB/year. Note, GTL information for a single microbe generates 20 PB of proteomic data and 16 PB of metabolite data.	Proteomics and metabolomics data generation has the potential to increase to the level of tens of PB/year by 2008.
Particle Physics In the search for the fundamental building blocks of the universe, the discovery of the top quark at the FNAL in 1995 required 70 TB of data and analysis from 1992 to 1995.	For the search for the Higgs boson at FNAL, 500 TB/year of data and analysis are currently being put to media.	Investigation of the properties of the Higgs boson will result in CERN Large Hadron Collider experiments acquiring 10 PB/year of data. 3-4 PB/year of the data will be moved to BNL and FNAL, and then onto U.S. universities, beginning in 2007. Processing this data will generate several additional PB/year.
Universe Asymmetry BaBar's mission at SLAC is to discover why our universe has an asymmetric distribution of matter and anti-matter. It went on-line in 1999.	BaBar currently has 200 TB/year of data and analysis going to media. To date, over a PB has been moved to partners in Europe for analysis.	Upgrades to the PEP-II accelerator will result in a quadrupling of BaBar's 2003 rate to close to 1 PB/year going to media as it searches for a deep understanding of processes at the origin of our universe.

As seen in the table above, on average from 1998 to 2008, there will be a 500- to 1,000-fold increase in the amount of data going to media at many DOE Office of Science facilities. As systems become more distributed and more integrated the amount of data transported on demand (as well as in an organized fashion) increases more rapidly than the amount of data acquired and processed at the central laboratories. Hence, 1,000 times per decade may be an underestimate, especially as effective data-intensive grid systems are built. These estimates roughly match the doubling seen every year in the amount of traffic moving across ESnet. What follows is a summary of the key factors driving this increase:

- The most important factor is that for many experiments, more data results in the increased potential for scientific discovery. In addition, the faster the data can be acquired, analyzed, and simulated, the faster the pace of scientific discovery. Scientists are very motivated to get data as rapidly as possible.
- Moore's Law of doubling the density of electronic circuits every 18 months applies to detectors as well as computers. Scientists have been very aggressive in increasing the spatial resolutions of their detectors. This corresponds to greatly increased channel density and consequently substantial increases in their data rates.
- For many scientific instruments, there are two additional dimensions that can increase data rates even faster than Moore's Law. The 100 megahertz clock speeds of the early 1990s have been replaced by gigahertz speeds in 2003 and will increase by close to a factor of 10 by 2008. This means that the ever higher density detectors are also pumping out data faster and faster. The second additional dimension is that for some experiments, the instruments can be layered in the physical third dimension. As the instruments shrink and their component costs decreases, multilayer instruments will become more common. Again, the result is data going to storage media at higher rates.
- Simulations have matured to the level that they are now considered to be the third leg of science, complementing theory and experiment. High-end computers have been growing in capabilities even faster than desktop computers. In terms of producing data from simulations, the software environments for high-end computers have advanced in capabilities at rates matching or exceeding Moore's Law. For many areas of science, high-end computers now generate and store simulation data to media at rates comparable to experiments, and in some cases exceed them.
- Experimental and/or simulation data stored in media (raw experimental data, analyzed data, simulated data, etc.) is typically analyzed by multiple scientists using multiple tools. Sometimes these tasks are carried out on very high-end visualization systems, but more often on a scientist's desktop. The capabilities of these desktop computers have been doubling roughly every 18 months. Since 1996, the disks for desktop computers have been increasing in storage density even faster, at rates over 100% per year. This rate is projected to return to the Moore's Law rate of 60% per year for the next 5 years. As seen in the table above, scientists have vast stores of data that they frequently move to and from their desktops and through multiple computational systems.

2.1 Science-Driven Collaboratories

A number of DOE large-scale science projects critically depend on collaborations of multidisciplinary researchers who collectively require capabilities that are unavailable at any single national laboratory or university. These projects span a wide spectrum of disciplines, including high-energy physics, climate simulation, fusion energy, genomics, and astrophysics, among others. In addition, the new experimental facilities coming on-line, such as ITER, LHC, and SNS, as well as the currently active facilities, such as ALS, APS, CEBAF, EMSL, FNAL Tevatron (Run II of CDF and D0), NLS, RHIC, and the SLAC PEP-II accelerator (BaBar), SSRL, and others, present unprecedented requirements for distributed, collaborative data analysis. These collaborations invariably involve geographically distributed resources such as supercomputers and clusters that offer massive computational speeds, user facilities that offer unique experimental capabilities, and repositories of experimental and computational data. These teams of researchers could be dispersed across the country or around the globe. Compounding the problem in some cases, access to these facilities must be tightly coordinated and controlled over wide-area networks. Indeed, seamless access to these distributed resources by researchers is essential to carrying out DOE missions, and the "network" and the associated collaboratory or grid tools have become critical components of the modern scientific infrastructure, much like the supercomputers or experimental facilities.

The DOE Office of Science envisions a seamless, high-performance network infrastructure to facilitate collaborations among researchers and their access to remote experimental and computational resources. Such an infrastructure can eliminate resource isolation, discourage redundancy, and promote rapid scientific progress through the interplay of theory, simulation, and experiment. For example, timely distribution of multi-petabytes of LHC data produced at CERN, in Switzerland, can eliminate the bottleneck experienced by U.S. physicists today due to inadequate bandwidth in the trans-Atlantic and U.S. networks. Also, the ability to remotely access complex scientific instruments in real time will enable interactive collaborations among geographically dispersed researchers, without the need for coordinated travel and duplications of specialized experimental instruments. An example is ITER, where it is envisaged that the new facility will be operated remotely by teams of geographically dispersed researchers from across the world.

In the August 2002 workshop, representatives of a range of DOE science disciplines were asked to provide information on how they currently use networking and network-associated services and what they saw as the future process of their science that would require, or be enabled by, adequate high-performance computers, high-speed networks, and advanced middleware support. Climate modeling has been picked as one of four examples from the August 2002 workshop to illustrate the importance of networks with enhanced services as part of an integrated cyber infrastructure for science.

Better climate modeling [4] is essential to understanding phenomena such as hurricanes, droughts and precipitation pattern changes, heat waves and cold snaps, and other potential changes that, e.g., promote disease-producing organisms or impact crop productivity. Better climate modeling requires very high-performance computing to permit simulation of realistic spatial and temporal resolution — it makes a huge difference in our ability to accommodate the impact of a sustained drought if we know the county-level geographic extent of the drought ten or twenty years in advance, rather than only that a drought is likely in this century and that it will affect the Midwest.

"Climate model" is a bit of a misnomer because the climate is determined by a complex interplay of physical and biological phenomena (See Figure 2-1). There are dozens of models connected by feedback loops that must be included in a realistic simulation of climate that will result in the accuracy needed to inform policy and advance planning issues that are critical for the well being of our society. The complexity of climate is typical of most macro-scale phenomena from cosmology to cellular function, so the issues raised by climate modeling are characteristic of much of science.

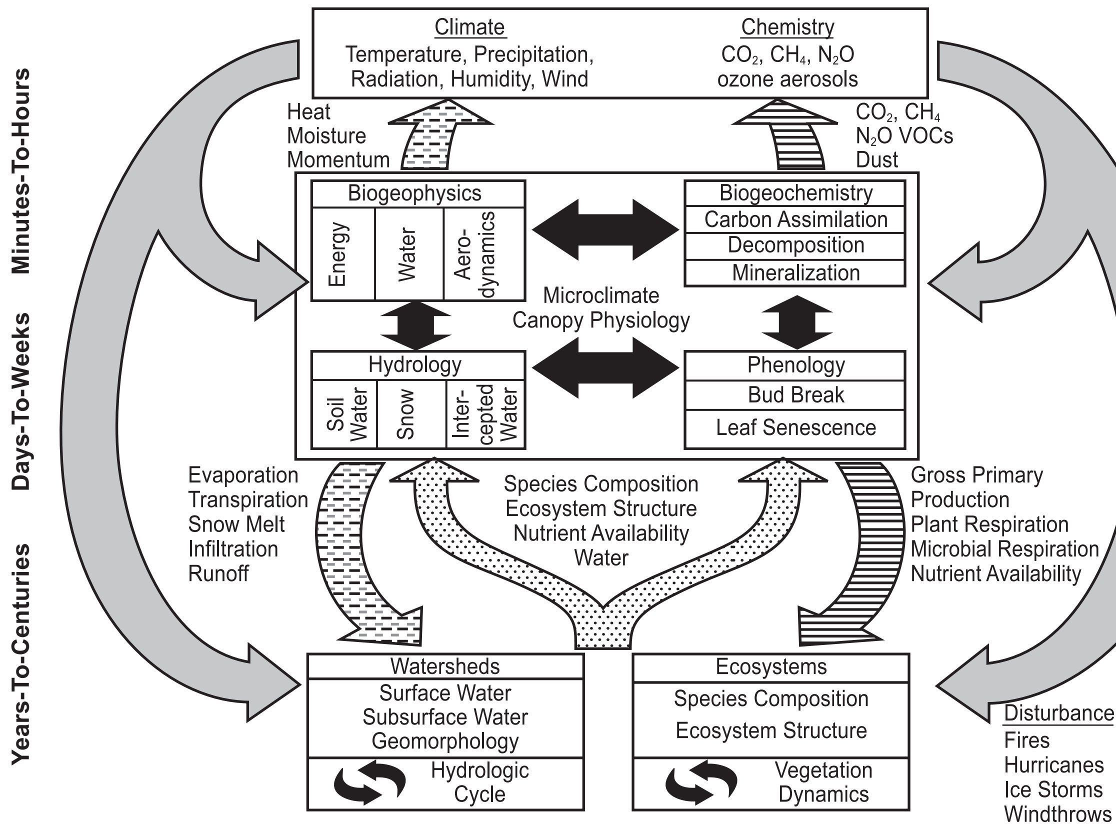


Figure 2-1 The Complexity of Climate Simulation [5]

Since the climate is an extremely complex phenomenon that involves modeling many separate elements in order to obtain the required accuracy, each of these elements is a discipline in its own right, and is studied by a different group of specialists.

Better climate modeling requires that the many institutions working on various aspects of climate be able to easily describe, catalogue, and seamlessly share the knowledge and the vast amounts of data that underlay the knowledge in order to facilitate the required interdisciplinary collaboration. Nonetheless, all of these sub-models must interoperate in the same way that all of the elements that make up the climate interact. This multidisciplinary simulation produces an inherently distributed computing environment as the models of the discipline's specialists are accessed and combined into an overall model of the climate via the collaboratory or grid environment.

Further, the many specialized scientific groups that work on the different components that go into a comprehensive model build specialized software and data environments that will probably never be homogenized and combined on a single computing system. Almost all multidisciplinary simulations are inherently distributed, with the overall simulation consisting of software and data on many different systems combined into a virtual system by using collaboratory tools and facilities for building distributed systems.
This, then, represents the vision of the future process of science in the climate community — to have the necessary computing power, access to annotated data, and interoperation of diverse sub-models, i.e., a collaboratory such that a realistic model of climate can make predictions that have great value to human society.

2.2 Science-Driven Evolution of Common Network Services

Over the last two decades, DOE science managers took several key steps called for by the rapid expansion of data and collaborations needed to achieve DOE science. In the mid-1980s, the utility of improved networking between DOE laboratories and their university collaborators was recognized, and several networks, including the High Energy Physics Network (HEPnet) and the Magnetic Fusion Energy Network (MFEnet), that ran different protocols were combined to form the ESnet. Although ESnet started as a multi-protocol network, the Internet Protocol (IP) is now used throughout because of its compatibility with the university communities and commercial vendor tools. Beginning in the early 1990s, the development of collaboratory tools began in earnest. Initially they were focused on distributed computing, file sharing, and instrument control. Three 1990s DOE/SC/ASCR/MICS programs in this area were the Distributed Informatics, Computing, & Collaborative Environment (DICCE); the Distributed Computing Experimental Environment (DCEE); and DOE2000 Collaboratories. The DOE2000 program is now expanding to include grid technologies, and currently the Scientific Discovery through Advanced Computing (SciDAC) program and MICS are supporting the R&D and implementation of grid-style collaboratory and computational tools for DOE science. The following table gives some examples of how the collaboratory tools are advancing in the DOE science environment.

While almost everyone connected to the Internet use tools, such as e-mail, it is largely the geographically distributed science community with its petabytes of data that is driving the usage of computational grids, remote instrument control, and collaborative visualizations, and DOE scientists with their vast research facilities are among those who are leading the way. To keep DOE science on track for the coming five years and longer, the networking and collaborative tools will need to match both the explosive growth of scientific data and the collaborative resources needed to produce, analyze, and understand the data. The R&D going into collaboratory tools and grid technologies will need to move into production services as long-term infrastructure available to support the mission of the Office of Science.

Table 2-2 DOE Science Community Use in Percent of Middleware Services

Middleware Service	1998	2003	2008
IP based audio-video/Access Grid/VRVS	<5	10	80
IP based telephone	<1	5	30
ISDN based video conferencing	10	50	5
Global directories – finding people	50	80	95
Global directories – finding services	<1	20	80
Computational grids for science	<1	20	80
Remote instrument control	<5	10	50
Collaboratively shared visualization	<1	10	50
Web services/portals	<2	20	80
Security infrastructure – PKI/certificates	<1	20	80
Security infrastructure – secure protocols	10	80	99

Due to the importance of the collaboratory tools, serious consideration has been given to changing the name of the Energy Sciences Network to something more inclusive. The argument for doing this is that a new name would symbolically capture the broader impact of ESnet and the collaboratory tools of DOE science. The argument against would be that in the networking community and across the Office of Science and in Congressional committees that support DOE, ESnet is recognized as one of the best (if not the best) networks in the world for support of science. In this report, we propose that the networking portion of the larger enterprise remain known as ESnet and that a new umbrella name such as Science Networking and Services be deployed to include both ESnet and the collaboratory/grid environment for DOE science.

In summary, as seen by the science drivers presented above, it is projected that Office of Science networking and services requirements will continue to double (or more) every year for the next five years (as they have since 1992). Meeting these networking requirements will require research and development specifically targeted at Office of Science networking issues. In addition, grid-style collaboratory tools will need the projected enhancements to be able to be used in efficiently and effectively managing the data and achieving the scientific discoveries that are the mission of the Office of Science.

2.3 References and Notes

1. Moore, G.E., *Electronics* 38(8), April 19, 1965. Available at: ftp://download.intel.com/research/silicon/moorespaper.pdf and http://www.intel.com/research/silicon/mooreslaw.htm.
2. International Technology Roadmap for Semiconductors 2002 Update. Available at: http://public.itrs.net.
3. See: http://public.web.cern.ch/public/about/achievements/www/www.html.
4. This scenario is due to Al Kellie, Gary Strand, and Tim Killeen of the National Center for Atmospheric Research (NCAR).
5. Figure courtesy of Gordon Bonan, NCAR. It is taken from Bonan, G., *Ecological Climatology: Concepts and Applications*, Cambridge: Cambridge University Press (2002).